

# Sparse Recovery with Orthogonal Matching Pursuit under RIP

Tong Zhang, *Member, IEEE*,

**Abstract**—This paper presents a new analysis for the orthogonal matching pursuit (OMP) algorithm. It is shown that if the restricted isometry property (RIP) is satisfied at sparsity level  $O(\bar{k})$ , then OMP can stably recover a  $\bar{k}$ -sparse signal in 2-norm under measurement noise. For compressed sensing applications, this result implies that in order to uniformly recover a  $\bar{k}$ -sparse signal in  $\mathbb{R}^d$ , only  $O(\bar{k} \ln d)$  random projections are needed. This analysis improves some earlier results on OMP depending on stronger conditions that can only be satisfied with  $\Omega(\bar{k}^2 \ln d)$  or  $\Omega(\bar{k}^{1.6} \ln d)$  random projections.

**Index Terms**—Estimation theory, feature selection, greedy algorithms, statistical learning, sparse recovery

## I. INTRODUCTION

Consider a signal  $\bar{\mathbf{x}} \in \mathbb{R}^d$ , and suppose that we observe its linear transformation plus measurement noise as:

$$\mathbf{y} = A\bar{\mathbf{x}} + \text{noise}.$$

Here,  $A$  is an  $n \times d$  matrix. If we define an objective function

$$Q(\mathbf{x}) = \|A\mathbf{x} - \mathbf{y}\|_2^2, \quad (1)$$

then we may estimate the parameter  $\bar{\mathbf{x}}$  by minimizing  $Q(\mathbf{x})$ , subject to appropriate constraints.

If  $d > n$ , then the solution of the unconstrained optimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^d} Q(\mathbf{x}) \quad (2)$$

is not unique. In order to estimate  $\bar{\mathbf{x}}$ , additional assumptions on  $\bar{\mathbf{x}}$  is necessary. We are specifically interested in the case where  $\bar{\mathbf{x}}$  is sparse. That is  $\|\bar{\mathbf{x}}\|_0 \ll n$ , where

$$\|\mathbf{x}\|_0 = |\text{supp}(\mathbf{x})|, \quad \text{supp}(\mathbf{x}) = \{j : x_j \neq 0\}.$$

It is known that under appropriate conditions, it is possible to recover  $\bar{\mathbf{x}}$  by solving (2) with a sparsity constraint as follows:

$$\min_{\mathbf{x} \in \mathbb{R}^d} Q(\mathbf{x}) \quad \text{subject to } \|\mathbf{x}\|_0 \leq k. \quad (3)$$

However, this optimization problem is generally NP-hard. Therefore one seeks computationally efficient algorithms that can approximately solve (3), with the goal of recovering sparse signal  $\bar{\mathbf{x}}$ . This paper considers the popular orthogonal matching pursuit algorithm (OMP), which has been widely used for this purpose (for example, see [5], [14], [15]). We are specifically interested in two issues: the performance of OMP in terms of optimizing  $Q(\mathbf{x})$  and the performance of OMP in terms of recovering the sparse signal  $\bar{\mathbf{x}}$ .

T. Zhang is with the Statistics Department, Rutgers University, New Jersey, USA. E-mail: tzhang@stat.rutgers.edu. The author was partially supported by the following grants: AFOSR-10097389, NSA-AMS 081024, NSF DMS-1007527, and NSF IIS-1016061.

## II. MAIN RESULT

Our analysis considers a more general objective function  $Q(\mathbf{x})$  that does not necessarily take the quadratic form in (1). However, we assume that  $Q(\mathbf{x})$  is convex. For such a general convex objective function, we consider the fully (or totally) corrective greedy algorithm in Figure 1, which was analyzed in [13]. This paper refines the analysis to show that the algorithm works under the restricted isometry property (RIP) of [3] (the required condition will be described later in this section). This algorithm is a direct generalization of OMP which has been traditionally considered only for the quadratic objective function in (1) with  $F^{(0)} = \emptyset$ . For simplicity, we assume that the number of iterations  $k_0$  is chosen a priori. The algorithm has been known in the machine learning community as a version of boosting [16], and has also been proposed recently in the signal processing community [2].

Input:  $Q(\mathbf{x})$  defined on  $\mathbb{R}^d$ ,  
initial feature set  $F^{(0)} \subset \{1, \dots, d\}$ .  
Output:  $\mathbf{x}^{(k)}$   
let  $\mathbf{x}^{(0)} = \arg \min_{\mathbf{x} \in \mathbb{R}^d} Q(\mathbf{x})$  subject to  $\text{supp}(\mathbf{x}) \subset F^{(0)}$   
(default choice is  $F^{(0)} = \emptyset$  with  $\mathbf{x}^{(0)} = 0$ )  
**for**  $k = 1, 2, \dots, k_0$   
    let  $j = \arg \max_i |\nabla Q(\mathbf{x}^{(k-1)})_i|$   
    let  $F^{(k)} = \{j\} \cup F^{(k-1)}$   
    let  $\mathbf{x}^{(k)} = \arg \min_{\mathbf{x} \in \mathbb{R}^d} Q(\mathbf{x})$  subject to  $\text{supp}(\mathbf{x}) \subset F^{(k)}$   
**end**

Fig. 1. Fully Corrective Greedy Boosting Algorithm (OMP)

For quadratic loss, the objective function  $Q(\mathbf{x})$  is given by (1) and its derivative is  $\nabla Q(\mathbf{x}) = 2A^\top(A\mathbf{x} - \mathbf{y})$ . Therefore  $j = \arg \max_i |\nabla Q(\mathbf{x}^{(k-1)})_i|$  becomes  $j = \arg \max_i |\mathbf{a}_i^\top(A\mathbf{x} - \mathbf{y})|$ , where  $\mathbf{a}_i$  is the  $i$ -th column of matrix  $A$ . This, together with  $F^{(0)} = \emptyset$ , leads to the standard OMP algorithm. In order to use notation consistent with the sparse recovery literature, in the current paper, we still refer to the more general algorithm in Figure 1 as OMP even though it applies to objective functions other than (1).

The general problem of optimization under sparsity constraint is NP hard. In order to alleviate the difficulty, we consider approximate optimization under the restricted strong convexity assumption introduced below.

**Definition 2.1 (Restricted Strong Convexity Constants):**  
Given any  $s \geq 0$ , define restricted strong convexity constants

$\rho_-(s)$  and  $\rho_+(s)$  as follows: for all  $\|\mathbf{x} - \mathbf{x}'\|_0 \leq s$ , we require

$$\begin{aligned} \rho_-(s)\|\mathbf{x} - \mathbf{x}'\|_2^2 &\leq Q(\mathbf{x}') - Q(\mathbf{x}) - \nabla Q(\mathbf{x})^\top (\mathbf{x}' - \mathbf{x}) \\ &\leq \rho_+(s)\|\mathbf{x} - \mathbf{x}'\|_2^2. \end{aligned}$$

If the objective function takes the quadratic form given by (1), then the above definition is equivalent to the following sparse eigenvalue condition of  $A^\top A$ :  $\forall \Delta \mathbf{x} \in \mathbb{R}^d$  such that  $\|\Delta \mathbf{x}\|_0 \leq s$ ,

$$\rho_-(s)\|\Delta \mathbf{x}\|_2^2 \leq \|A\Delta \mathbf{x}\|_2^2 \leq \rho_+(s)\|\Delta \mathbf{x}\|_2^2. \quad (4)$$

In this case, the constants  $\rho_-(s)$  and  $\rho_+(s)$  are closely related to the restricted isometry constant  $\delta_s$  in [3], which is defined as a constant that satisfies the condition that  $\forall \Delta \mathbf{x} \in \mathbb{R}^d$  such that  $\|\Delta \mathbf{x}\|_0 \leq s$ :

$$(1 - \delta_s)\|\Delta \mathbf{x}\|_2^2 \leq \|A\Delta \mathbf{x}\|_2^2 \leq (1 + \delta_s)\|\Delta \mathbf{x}\|_2^2.$$

The restricted isometry constant was used to define the restricted isometry property (RIP) in the analysis of  $L_1$  regularization method [3]. We employ the slightly more general restricted strong convexity constants in (4) because our analysis only requires the ratio  $\rho_+(s)/\rho_-(s)$  to be bounded, and this is useful for general machine learning problems where  $\rho_+(s)$  can be larger than 2.

In order to recover the target  $\bar{\mathbf{x}}$ , we have to assume that  $\bar{\mathbf{x}}$  is sparse and approximately optimizes  $Q(\mathbf{x})$ . If a target  $\bar{\mathbf{x}}$  is an exact global optimal solution, then  $\nabla Q(\bar{\mathbf{x}}) = 0$ . However, this paper deals with approximate optimal solutions, where  $\nabla Q(\bar{\mathbf{x}}) \approx 0$ . In particular, we introduce the following definition, which is convenient to apply.

*Definition 2.2 (Restricted Gradient Optimal Constant):*

Given  $\bar{\mathbf{x}} \in \mathbb{R}^d$  and  $s > 0$ , we define the restricted gradient optimal constant  $\epsilon_s(\bar{\mathbf{x}})$  as the smallest non-negative value that satisfies the following condition

$$|\nabla Q(\bar{\mathbf{x}})^\top \mathbf{u}| \leq \epsilon_s(\bar{\mathbf{x}})\|\mathbf{u}\|_2$$

for all  $\mathbf{u} \in \mathbb{R}^d$  such that  $\|\mathbf{u}\|_0 \leq s$ .

The constant  $\epsilon_s(\bar{\mathbf{x}})$  measures how close is  $\nabla Q(\bar{\mathbf{x}})$  to zero. If  $\nabla Q(\bar{\mathbf{x}}) = 0$ , then  $\epsilon_s(\bar{\mathbf{x}}) = 0$ . If  $\nabla Q(\bar{\mathbf{x}}) \approx 0$ , then  $\epsilon_s(\bar{\mathbf{x}})$  is small. Moreover, similar to the definition of restricted strong convex constants, we are only interested in the value of  $\nabla Q(\bar{\mathbf{x}})$  in any subset of  $\{1, \dots, d\}$  with  $s$  elements. The following proposition provides some estimates of  $\epsilon_s(\bar{\mathbf{x}})$  using quantities that are easier to understand.

*Proposition 2.1:* We have  $\epsilon_s(\bar{\mathbf{x}}) \leq \sqrt{s}\|\nabla Q(\bar{\mathbf{x}})\|_\infty$  and  $\epsilon_s(\bar{\mathbf{x}}) \leq \|\nabla Q(\bar{\mathbf{x}})\|_2$ . Moreover, if

$$Q(\bar{\mathbf{x}}) \leq \inf_{\|\mathbf{x}\|_0 \leq \|\bar{\mathbf{x}}\|_0 + s} Q(\mathbf{x}) + \bar{\epsilon},$$

then

$$\epsilon_s(\bar{\mathbf{x}}) \leq 2\sqrt{\rho_+(s)\bar{\epsilon}}.$$

*Proof:* The first two inequalities are straight-forward. For the third inequality, we note that for  $\|\mathbf{u}\|_0 \leq s$ :

$$\begin{aligned} &\inf_{\|\mathbf{x}\|_0 \leq \|\bar{\mathbf{x}}\|_0 + s} Q(\mathbf{x}) \\ &\leq \inf_{\eta} Q(\bar{\mathbf{x}} + \eta \mathbf{u}) \\ &\leq \inf_{\eta} [Q(\bar{\mathbf{x}}) + \eta \nabla Q(\bar{\mathbf{x}})^\top \mathbf{u} + \rho_+(s)\eta^2\|\mathbf{u}\|_2^2] \\ &= Q(\bar{\mathbf{x}}) - |\nabla Q(\bar{\mathbf{x}})^\top \mathbf{u}|^2 / (4\rho_+(s)\|\mathbf{u}\|_2^2). \end{aligned}$$

The result follows by rearranging the above inequality.  $\blacksquare$

The following theorem is the main result of this paper, which shows that OMP can approximately recover a sparse signal  $\bar{\mathbf{x}}$  in 2-norm if the condition (5) in the theorem involving strong convexity constants can be satisfied. As we shall discuss later, this condition is closely related to the RIP condition for the quadratic objective (1).

*Theorem 2.1:* Consider the OMP algorithm. Let  $\bar{\mathbf{x}} \in \mathbb{R}^d$  and  $\bar{F} = \text{supp}(\bar{\mathbf{x}})$ . If there exists  $s$  such that

$$\begin{aligned} s &\geq |\bar{F} \cup F^{(0)}| \\ &\quad + 4|\bar{F} \setminus F^{(0)}| \frac{\rho_+(1)}{\rho_-(s)} \ln \frac{20\rho_+(|\bar{F} \setminus F^{(0)}|)}{\rho_-(s)}, \end{aligned} \quad (5)$$

then when  $k = k_0 = s - |\bar{F} \cup F^{(0)}|$ , we have

$$Q(\mathbf{x}^{(k)}) \leq Q(\bar{\mathbf{x}}) + 2.5\epsilon_s(\bar{\mathbf{x}})^2 / \rho_-(s)$$

and

$$\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\|_2 \leq \sqrt{6}\epsilon_s(\bar{\mathbf{x}}) / \rho_-(s).$$

*Proof:* The detailed proof relies on a number of technical lemmas that are left to the appendix.

The first inequality of the theorem is a direct consequence of Lemma A.5. The second inequality is a consequence of the first inequality and Lemma A.2:

$$\begin{aligned} &\rho_-(s)\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\|_2^2 \\ &\leq 2 \left[ Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}}) \right] + \epsilon_s(\bar{\mathbf{x}})^2 / \rho_-(s) \\ &\leq 6\epsilon_s(\bar{\mathbf{x}})^2 / \rho_-(s). \end{aligned}$$

This implies the second inequality.  $\blacksquare$

Note that (5) can be satisfied as long as  $(\rho_+(1)/\rho_-(s)) \ln(\rho_+(|\bar{k}|)/\rho_-(s))$  grows sub-linearly as a function of  $s$ . With appropriate assumptions, this allows the ratio  $\rho_+(s)/\rho_-(s)$  to be significantly larger than 1 but bounded from above (such a condition is sometimes referred to as sparse eigenvalue condition in the statistics literature). In this context, Theorem 2.1 is useful for estimation problems encountered in machine learning, where  $\rho_+(s)/\rho_-(s)$  may be large.

In compressed sensing, one can often control the ratio of  $\rho_+(s)/\rho_-(s)$  to be not much larger than 1 using random projection. In this context, the following result gives a simpler interpretation of the above theorem, where the condition (5) of the theorem is replaced by  $\rho_+(\bar{k}) \leq 2\rho_-(31\bar{k})$ .

*Corollary 2.1:* Consider the OMP algorithm with  $F^{(0)} = \emptyset$ . Let  $\bar{\mathbf{x}} \in \mathbb{R}^d$  and  $\bar{k} = \|\bar{\mathbf{x}}\|_0$ . If the condition  $\rho_+(\bar{k}) \leq 2\rho_-(31\bar{k})$  holds, then when  $k = k_0 = 30\bar{k}$ , we have

$$Q(\mathbf{x}^{(k)}) \leq Q(\bar{\mathbf{x}}) + 2.5\epsilon_s(\bar{\mathbf{x}})^2 / \rho_-(s)$$

and

$$\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\|_2 \leq \sqrt{6}\epsilon_s(\bar{\mathbf{x}})/\rho_-(s),$$

where  $s = 31\bar{k}$ .

*Proof:* If  $\rho_+(\bar{k}) \leq 2\rho_-(31\bar{k})$  holds, then we can let  $s = 31\bar{k}$ , which implies that

$$2 \geq \rho_+(\bar{k})/\rho_-(s) \geq \rho_+(1)/\rho_-(s).$$

Therefore

$$\begin{aligned} s &= 30\bar{k} \geq \bar{k} + 4\bar{k} \cdot 2 \ln(20 \cdot 2) \\ &\geq \bar{k} + 4\bar{k}(\rho_+(1)/\rho_-(s)) \ln(20\rho_+(\bar{k})/\rho_-(s)). \end{aligned}$$

This means that the condition (5) holds, and the corollary follows directly from Theorem 2.1.  $\blacksquare$

For the quadratic objective (1), the condition  $\rho_+(\bar{k}) \leq 2\rho_-(31\bar{k})$  is analogous to the RIP condition in [3]. In particular, if the matrix  $A$  has the restricted isometry constant  $\delta_{31\bar{k}} \leq 1/3$ , then the condition  $\rho_+(\bar{k}) \leq 4/3$  and  $\rho_-(31\bar{k}) \geq 2/3$  holds, with  $\rho_+(s)$  and  $\rho_-(s)$  defined according to (4). In this case, Corollary 2.1 can be directly applied.

It is interesting to observe that except for constants, the result of this paper for OMP is as strong as those for more sophisticated greedy algorithms such as ROMP [11] or CoSaMP [10]. For example, Corollary 2.1 can be applied when  $\delta_s \leq 1/3$  with  $s = 31\bar{k}$ , while a similar result for CoSaMP in [10] applies when  $\delta_s \leq 0.1$  with  $s = 4\bar{k}$ . Nevertheless, the difference in the constants may still suggest possible advantages for more complex algorithms such as CoSaMP under suitable conditions.

For quadratic objective function, a simple instantiation of  $\epsilon_s(\bar{\mathbf{x}})$  using Proposition 2.1 leads to the following sparse recovery result that is relatively simple to interpret.

*Corollary 2.2:* Consider the quadratic objective function  $Q(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$  of (1), and the OMP algorithm with  $F^{(0)} = \emptyset$ . Consider an arbitrary vector  $\bar{\mathbf{x}} \in \mathbb{R}^d$  and let  $\bar{k} = \|\bar{\mathbf{x}}\|_0$ . If the RIP condition  $\rho_+(\bar{k}) \leq 2\rho_-(31\bar{k})$  holds, then when  $k = k_0 = 30\bar{k}$ , we have

$$\|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\|_2 \leq 2\sqrt{6}\rho_+(s)^{1/2}\|\mathbf{A}\bar{\mathbf{x}} - \mathbf{y}\|_2/\rho_-(s),$$

where  $s = 31\bar{k}$ .

### III. DISCUSSION

In this paper we proved a new result for a generalization of the OMP algorithm. It is shown that if the RIP is satisfied at sparsity level  $O(\bar{k})$ , then OMP can recover a  $\bar{k}$ -sparse signal in 2-norm. For compressed sensing applications, this result implies that in order to uniformly recover a  $\bar{k}$ -sparse signal in  $\mathbb{R}^d$ , only  $n = O(\bar{k} \ln d)$  random projections are needed [3].

Our result for signal recovery is stronger than previous results for OMP that relied on different conditions. For example, [14] considered the problem of recovering the support set of a sparse signal under a stronger condition (also see [18] for recovery properties under stochastic noise). A similar analysis was employed in [15], where it was shown that for any fixed sparse signal  $\bar{\mathbf{x}}$  with  $\bar{k} = \|\bar{\mathbf{x}}\|_0$ , OMP can recover the signal with large probability using  $O(\bar{k} \ln d)$  measurements. A more refined analysis in [6] shows that a lower bound

of  $n = 2\bar{k} \ln(d - \bar{k})$  measurements is enough for recovery. However, the above results are not uniform with respect to all  $\bar{k}$ -sparse signals  $\bar{\mathbf{x}}$  (that is, for any set of random projections, there exist  $\bar{k}$ -sparsity signals that fail the analysis). In comparison, the RIP condition holds uniformly by definition, and hence our result applies uniformly to all  $\bar{k}$ -sparse signals. Although our result is stronger than previous results in terms of signal recovery in 2-norm, the result requires running the OMP algorithm for more than  $\bar{k}$  iterations, and hence doesn't recover the true support set of the ideal signal. In comparison, results such as [15] also imply exact recovery of the correct support set (but under stronger assumptions) using only  $\bar{k}$  OMP iterations. It is also known that it is impossible to uniformly recover the support set (in  $\bar{k}$  iterations) with the OMP algorithm with  $O(\bar{k} \ln d)$  measurements [12]. This means that it is necessary to run OMP for more than  $\bar{k}$  iterations in order to achieve the best 2-norm recovery performance with as few measurements as possible.

It is worth mentioning that some previous results apply uniformly to all  $\bar{k}$ -sparse signals. For example, results in [5] depend on the stronger mutual incoherence condition. Unfortunately the mutual incoherence condition can only be satisfied with  $\Omega(\bar{k}^2 \ln d)$  random projections. Therefore in recent years there have been significant interests in studying OMP under the RIP. In addition to the current paper, a number of recent papers investigated this issue, reaching varying conclusions [1], [4], [8], [9]. For example, the RIP-based analysis for sparse signals (but without noise) was considered in [4], [8], with the conclusion that under a sufficiently strong assumption on the RIP constant (in fact, the resulting condition is similar to the mutual incoherence condition), exact recovery is possible in  $\bar{k}$  iterations. The condition required for the RIP constant was weakened in [9], where the author showed that by running the OMP algorithm more than  $\bar{k}$  iterations, it is possible to achieve exact recovery (again assuming no noise). The condition in [9] can be satisfied with only  $O(\bar{k}^{1.6} \ln d)$  measurements, which is a significant improvement over the traditional  $\Omega(\bar{k}^2 \ln d)$  measurements. The result obtained in the current paper is along the same line as [9], but reduced the required number of measurements to the optimal order of  $O(\bar{k} \ln d)$ .

It is also interesting to compare the new OMP result in this paper to that of Lasso, which is also known to work under the RIP. However, a more refined comparison illustrates differences between the known theoretical results for these two methods. For OMP, the result in Theorem 2.1 can be applied as long as the condition

$$\begin{aligned} s/|\bar{F} \cup F^{(0)}| &\geq \\ 4|\bar{F} \setminus F^{(0)}|(\rho_+(1)/\rho_-(s)) \ln(20\rho_+(|\bar{F} \setminus F^{(0)}|)/\rho_-(s)) \end{aligned}$$

is satisfied. With  $F^{(0)} = \emptyset$ , this roughly requires  $(\rho_+(1)/\rho_-(s)) \ln(\rho_+(\bar{k})/\rho_-(s))$  to grow sub-linearly as a function of  $s$  in order to apply the theory. In comparison, the known condition for Lasso (e.g., this has been made explicit in [17], [19]) requires  $\rho_+(s)/\rho_-(s)$  to grow sub-linearly as a function of  $s$ . To compare the two conditions, we note that the condition for OMP is weaker in terms of of the

upper convexity constant as there is no explicit dependency on  $\rho_+(s)$ ; however, the dependency on  $\rho_-(s)$  is stronger in OMP than Lasso due to the logarithmic term. Although it is unclear how tight these conditions are, the comparison nevertheless indicates that even though both algorithms work under the RIP, there are still finer differences in their theoretical analysis: Lasso is slightly more favorable in terms of its dependency on the lower strong convexity constant, while OMP is more favorable in terms of its dependency on the upper strong convexity constant. We further conjecture that the extra logarithmic dependency  $\ln(\rho_+(\bar{k})/\rho_-(s))$  in OMP is necessary. In practice, some times Lasso performs better while other times OMP performs better (for example, see experimental results in [7]). Therefore some discrepancy in their theoretical analysis is expected. More specifically, for sparse recovery, one often observes that Lasso is superior when the nonzero coefficients have a similar magnitude (which happens to be the case that the extra  $\ln(\rho_+(\bar{k})/\rho_-(s))$  factor is required in our OMP analysis) while OMP performs better when the nonzero coefficients exhibit rapid decay (which happens to be the case that the extra  $\ln(\rho_+(\bar{k})/\rho_-(s))$  factor can be removed from our analysis). The theory in this paper significantly narrows the previous theoretical gap between these two sparse recovery methods by positively answering the open question of whether OMP can recover sparse signals under the RIP. Therefore our result allows practitioners to apply OMP with more confidence than previously expected.

#### ACKNOWLEDGEMENTS

The author would like to thank the anonymous referees for pointing out many relevant references and for suggestions to improve the presentation.

#### APPENDIX

We need a number of technical lemmas. Lemma A.3 and Lemma A.4, key to the proof, are based on earlier work of the author with collaborators [13], [7]. The first three lemmas use the following notations. Let  $F, \bar{F}$  be two subsets of  $\{1, \dots, d\}$ . Let  $\text{supp}(\bar{\mathbf{x}}) \subset \bar{F}$ , and

$$\mathbf{x} = \arg \min_{\mathbf{z}: \text{supp}(\mathbf{z}) \subset F} Q(\mathbf{z}).$$

*Lemma A.1:* We have

$$Q(\mathbf{x}) - Q(\bar{\mathbf{x}}) \leq 1.5\rho_+(s)\|\bar{\mathbf{x}}_{\bar{F} \setminus F}\|_2^2 + 0.5\epsilon_s(\bar{\mathbf{x}})^2/\rho_+(s)$$

for all  $s \geq |\bar{F} \setminus F|$ .

*Proof:* Let  $\mathbf{x}' = \bar{\mathbf{x}}_{\bar{F} \cap F}$ , then by the definition of  $\mathbf{x}$ , we know that  $Q(\mathbf{x}) \leq Q(\mathbf{x}')$ . Therefore

$$\begin{aligned} & Q(\mathbf{x}) - Q(\bar{\mathbf{x}}) \\ & \leq Q(\mathbf{x}') - Q(\bar{\mathbf{x}}) \\ & = Q(\mathbf{x}') - Q(\bar{\mathbf{x}}) - \nabla Q(\bar{\mathbf{x}})^\top (\mathbf{x}' - \bar{\mathbf{x}}) + \nabla Q(\bar{\mathbf{x}})^\top (\mathbf{x}' - \bar{\mathbf{x}}) \\ & \leq \rho_+(s)\|\bar{\mathbf{x}}_{\bar{F} \setminus F}\|_2^2 + \epsilon_s(\bar{\mathbf{x}})\|\bar{\mathbf{x}}_{\bar{F} \setminus F}\|_2 \\ & \leq \rho_+(s)\|\bar{\mathbf{x}}_{\bar{F} \setminus F}\|_2^2 + 0.5\epsilon_s(\bar{\mathbf{x}})^2/\rho_+(s) + 0.5\rho_+(s)\|\bar{\mathbf{x}}_{\bar{F} \setminus F}\|_2^2, \end{aligned}$$

which implies the lemma. The first inequality is by the definitions of  $\rho_+(s)$  and  $\epsilon_s(\bar{\mathbf{x}})$ . The last inequality follows

from the fact that  $ab \leq 0.5a^2 + 0.5b^2$  with  $a = \epsilon_s(\bar{\mathbf{x}})/\sqrt{\rho_+(s)}$  and  $b = \sqrt{\rho_+(s)}\|\bar{\mathbf{x}}_{\bar{F} \setminus F}\|_2$ . ■

*Lemma A.2:* We have:

$$\rho_-(s)\|\mathbf{x} - \bar{\mathbf{x}}\|_2^2 \leq 2[Q(\mathbf{x}) - Q(\bar{\mathbf{x}})] + \epsilon_s(\bar{\mathbf{x}})^2/\rho_-(s)$$

for all  $s \geq |F \cup \bar{F}|$ .

*Proof:* From

$$\begin{aligned} & Q(\mathbf{x}) - Q(\bar{\mathbf{x}}) \\ & = Q(\mathbf{x}) - Q(\bar{\mathbf{x}}) - \nabla Q(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) + \nabla Q(\bar{\mathbf{x}})^\top (\mathbf{x} - \bar{\mathbf{x}}) \\ & \geq \rho_-(s)\|\bar{\mathbf{x}} - \mathbf{x}\|_2^2 - \epsilon_s(\bar{\mathbf{x}})\|\bar{\mathbf{x}} - \mathbf{x}\|_2 \\ & \geq 0.5\rho_-(s)\|\bar{\mathbf{x}} - \mathbf{x}\|_2^2 - 0.5\epsilon_s(\bar{\mathbf{x}})^2/\rho_-(s), \end{aligned}$$

we obtain the desired inequality. The first inequality is by the definitions of  $\rho_-(s)$  and  $\epsilon_s(\bar{\mathbf{x}})$ . The last inequality again follows from the fact that  $ab \leq 0.5a^2 + 0.5b^2$  with  $a = \epsilon_s(\bar{\mathbf{x}})/\sqrt{\rho_+(s)}$  and  $b = \sqrt{\rho_+(s)}\|\bar{\mathbf{x}}_{\bar{F} \setminus F}\|_2$ . ■

The next lemma shows that each greedy search makes reasonable progress. This proof is essentially identical to a similar result in [13] but with refined notations used in the current paper. We thus include the proof for completeness. It allows the readers to verify more easily that the proof in [13] remains unchanged with our new definitions.

*Lemma A.3:* Let  $\mathbf{e}_i \in \mathbb{R}^d$  be the vector of zeros except for the  $i$ -th component being one. If  $\bar{F} \setminus F \neq \emptyset$ , then for all  $s \geq |F \cup \bar{F}|$ :

$$\begin{aligned} & \min_{\alpha} Q(\mathbf{x} + \alpha \mathbf{e}_j) \\ & \leq Q(\mathbf{x}) - \frac{\rho_-(s)\|\mathbf{x} - \bar{\mathbf{x}}\|_2^2}{\rho_+(1) \left( \sum_{i \in \bar{F} \setminus F} |\bar{\mathbf{x}}_i| \right)^2} \max(0, Q(\mathbf{x}) - Q(\bar{\mathbf{x}})), \end{aligned}$$

where  $j = \arg \max_i |\nabla Q(\mathbf{x})_i|$ .

*Proof:* For all  $i \in \{1, \dots, d\}$  and  $\eta > 0$ , we define

$$Q_i(\eta) = Q(\mathbf{x}) + \eta \text{sgn}(\bar{\mathbf{x}}_i) \nabla Q(\mathbf{x})_i + \eta^2 \rho_+(1).$$

It follows from the definition of  $\rho_+(1)$  that

$$\min_{\alpha} Q(\mathbf{x} + \alpha \mathbf{e}_j) \leq Q(\mathbf{x} + \eta \text{sgn}(\bar{\mathbf{x}}_j) \mathbf{e}_j) \leq Q_j(\eta).$$

Since the choice of  $j = \arg \max_i |\nabla Q(\mathbf{x})_i|$  achieves the minimum of  $\min_i \min_{\eta} Q_i(\eta)$ , the lemma is a direct consequence of the following stronger statement:

$$\begin{aligned} & \min_i Q_i(\eta) \\ & \leq Q(\mathbf{x}) - \frac{\max(0, Q(\mathbf{x}) - Q(\bar{\mathbf{x}}) + \rho_-(s)\|\mathbf{x} - \bar{\mathbf{x}}\|_2^2)^2}{4\rho_+(1) \left( \sum_{i \in \bar{F} \setminus F} |\bar{\mathbf{x}}_i| \right)^2}, \end{aligned} \tag{6}$$

with an appropriate choice of  $\eta$ ; this is because

$$\begin{aligned} & \max(0, Q(\mathbf{x}) - Q(\bar{\mathbf{x}}) + \rho_-(s)\|\mathbf{x} - \bar{\mathbf{x}}\|_2^2)^2 \\ & \geq 4\rho_-(s) \max(0, Q(\mathbf{x}) - Q(\bar{\mathbf{x}}))\|\mathbf{x} - \bar{\mathbf{x}}\|_2^2. \end{aligned}$$

Therefore, we now turn to prove that (6) holds. Denoting  $u = \sum_{i \in \bar{F} \setminus F} |\bar{\mathbf{x}}_i|$ , we obtain that

$$\begin{aligned} & u \min_i Q_i(\eta) \leq \sum_{i \in \bar{F} \setminus F} |\bar{\mathbf{x}}_i| Q_i(\eta) \\ & \leq u Q(\mathbf{x}) + \eta \sum_{i \in \bar{F} \setminus F} \bar{\mathbf{x}}_i \nabla Q(\mathbf{x})_i + u \rho_+(1) \eta^2. \end{aligned} \tag{7}$$

Since we assume that  $\mathbf{x}$  is optimal over  $F$ , we get that  $\nabla Q(\mathbf{x})_i = 0$  for all  $i \in F$ . Additionally,  $\mathbf{x}_i = 0$  for  $i \notin F$  and  $\bar{\mathbf{x}}_i = 0$  for  $i \notin \bar{F}$ . Therefore,

$$\begin{aligned} \sum_{i \in \bar{F} \setminus F} \bar{\mathbf{x}}_i \nabla Q(\mathbf{x})_i &= \sum_{i \in \bar{F} \setminus F} (\bar{\mathbf{x}}_i - \mathbf{x}_i) \nabla Q(\mathbf{x})_i \\ &= \sum_{i \in \bar{F} \cup F} (\bar{\mathbf{x}}_i - \mathbf{x}_i) \nabla Q(\mathbf{x})_i \\ &= \nabla Q(\mathbf{x})^\top (\bar{\mathbf{x}} - \mathbf{x}). \end{aligned}$$

Combining the above with the definition of  $\rho_-(s)$ , we obtain that

$$\sum_{i \in \bar{F} \setminus F} \bar{\mathbf{x}}_i \nabla Q(\mathbf{x})_i \leq Q(\bar{\mathbf{x}}) - Q(\mathbf{x}) - \rho_-(s) \|\mathbf{x} - \bar{\mathbf{x}}\|_2^2.$$

Combining the above with (7) we get

$$\begin{aligned} &u \min_i Q_i(\eta) \\ &\leq u Q(\mathbf{x}) + \eta [Q(\bar{\mathbf{x}}) - Q(\mathbf{x}) - \rho_-(s) \|\mathbf{x} - \bar{\mathbf{x}}\|_2^2] + u \rho_+(1) \eta^2. \end{aligned}$$

Setting

$$\eta = \max[0, Q(\mathbf{x}) - Q(\bar{\mathbf{x}}) + \rho_-(s) \|\mathbf{x} - \bar{\mathbf{x}}\|_2^2] / (2u \rho_+(1))$$

and rearranging the terms, we conclude our proof of (6). ■

The direct consequence of the previous lemma is the following result, which is critical in our analysis. The idea of using a nesting approximating sequence has appeared in [7], but the current version is improved. The change is necessary for the purpose of this paper. In the following  $\mu$  can be chosen as any positive number if  $L = 1$ .

*Lemma A.4:* Consider the OMP algorithm. Consider a positive integer  $L$  and subsets  $\bar{F}_0 \subset \bar{F}_1 \subset \bar{F}_2 \cdots \bar{F}_L \subset \bar{F} \cup F^{(0)}$ , where  $\bar{F}_0 = \bar{F} \cap F^{(0)}$ . Assume that  $\min_{\mathbf{x}: \text{supp}(\mathbf{x}) \subset \bar{F}_j} Q(\mathbf{x}) \leq Q(\bar{\mathbf{x}}) + q_j$  ( $j = 0, \dots, L$ ),  $q_0 \geq q_1 \geq \dots \geq q_L \geq 0$ , and let  $\mu \geq \sup_{j=1, \dots, L-1} (q_{j-1}/q_j)$ . If  $s \geq |F^{(k)} \cup \bar{F}|$  and

$$k = \sum_{j=1}^L \left[ |\bar{F}_j \setminus F^{(0)}| (\rho_+(1)/\rho_-(s)) \ln(2\mu) \right],$$

then

$$Q(\mathbf{x}^{(k)}) \leq Q(\bar{\mathbf{x}}) + q_L + \mu^{-1} q_{L-1}.$$

*Proof:* Note that for any  $\text{supp}(\mathbf{x}) \subset F$  and  $\text{supp}(\bar{\mathbf{x}}) \subset \bar{F}$ , we have when  $\bar{F} \setminus F \neq \emptyset$ :

$$\frac{\rho_-(s) \|\mathbf{x} - \bar{\mathbf{x}}\|^2}{\rho_+(1) \left( \sum_{i \in \bar{F} \setminus F} |\bar{\mathbf{x}}_i| \right)^2} \geq \frac{\rho_-(s)}{\rho_+(1) |\bar{F} \setminus F|}.$$

Therefore Lemma A.3 implies that at any  $k$  such that  $s \geq |F^{(k)} \cup \bar{F}|$  and  $\ell = 0, \dots, L$ , we have either  $|\bar{F}_\ell \setminus F^{(k)}| = 0$  or

$$\begin{aligned} Q(\mathbf{x}^{(k+1)}) &\leq Q(\mathbf{x}^{(k)}) - \\ &\quad \frac{\rho_-(s)}{\rho_+(1) |\bar{F}_\ell \setminus F^{(k)}|} \max \left( 0, Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}}) - q_\ell \right), \end{aligned}$$

where we simply replace the target vector  $\bar{\mathbf{x}}$  in Lemma A.3 by the optimal solution over  $\bar{F}_\ell$ , and replace  $\mathbf{x}$  by  $\mathbf{x}^{(k)}$ . The

inequality, along with  $Q(\mathbf{x}^{(k+1)}) \leq Q(\mathbf{x}^{(k)})$ , implies that either  $|\bar{F}_\ell \setminus F^{(k)}| = 0$  or

$$\begin{aligned} &\max(0, Q(\mathbf{x}^{(k+1)}) - Q(\bar{\mathbf{x}}) - q_\ell) \\ &\leq \left[ 1 - \frac{\rho_-(s)}{\rho_+(1) |\bar{F}_\ell \setminus F^{(k)}|} \right] \max \left( 0, Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}}) - q_\ell \right) \\ &\leq \exp \left[ -\frac{\rho_-(s)}{\rho_+(1) |\bar{F}_\ell \setminus F^{(k)}|} \right] \max \left( 0, Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}}) - q_\ell \right). \end{aligned}$$

Therefore for any  $k' \leq k$  and  $\ell = 1, \dots, L$ , we have either  $|\bar{F}_\ell \setminus F^{(k)}| = 0$  or

$$\begin{aligned} &Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}}) - q_\ell \leq \\ &\exp \left[ -\frac{\rho_-(s)(k - k')}{\rho_+(1) |\bar{F}_\ell \setminus F^{(k')}|} \right] \max \left( 0, Q(\mathbf{x}^{(k')}) - Q(\bar{\mathbf{x}}) - q_\ell \right). \end{aligned} \quad (8)$$

We are now ready to prove the lemma by induction on  $L$ . If  $L = 1$ , we can set  $k' = 0$  in (8) and consider any  $\mu > 0$ . Since  $Q(\mathbf{x}^{(0)}) \leq \min_{\mathbf{x}: \text{supp}(\mathbf{x}) \subset \bar{F}_0} Q(\mathbf{x}) \leq Q(\bar{\mathbf{x}}) + q_0$ , we have

$$Q(\mathbf{x}^{(0)}) - Q(\bar{\mathbf{x}}) - q_1 \leq q_0.$$

Therefore when

$$k = \left\lceil |\bar{F}_1 \setminus F^{(0)}| (\rho_+(1)/\rho_-(s)) \ln(2\mu) \right\rceil,$$

we have from (8) that if  $|\bar{F}_1 \setminus F^{(k)}| \neq 0$ , then

$$\begin{aligned} &Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}}) - q_1 \\ &\leq \exp \left[ -\frac{\rho_-(s)k}{\rho_+(1) |\bar{F}_1 \setminus F^{(0)}|} \right] q_0 \\ &\leq (2\mu)^{-1} q_0. \end{aligned}$$

Note that this inequality also holds when  $|\bar{F}_1 \setminus F^{(k)}| = 0$ , and in such case (8) does not apply. This is because in this case  $Q(\mathbf{x}^{(k)}) \leq \min_{\mathbf{x}: \text{supp}(\mathbf{x}) \subset \bar{F}_1} Q(\mathbf{x}) \leq Q(\bar{\mathbf{x}}) + q_1$ . Therefore the lemma always holds when  $L = 1$ .

Now assume that the lemma holds at  $L = m - 1$  for some  $m > 1$ . That is, with

$$k' = \sum_{j=1}^{m-1} \left\lceil |\bar{F}_j \setminus F^{(0)}| (\rho_+(1)/\rho_-(s)) \ln(2\mu) \right\rceil,$$

we have

$$Q(\mathbf{x}^{(k')}) \leq Q(\bar{\mathbf{x}}) + q_{m-1} + \mu^{-1} q_{m-2}.$$

This implies that when  $L = m$ :

$$Q(\mathbf{x}^{(k')}) - Q(\bar{\mathbf{x}}) - q_L \leq q_{L-1} + \mu^{-1} q_{L-2} - q_L \leq 2q_{L-1}.$$

We thus obtain from (8) that if  $|\bar{F}_L \setminus F^{(k)}| \neq 0$ , then

$$\begin{aligned} &Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}}) - q_L \\ &\leq \exp \left[ -\frac{\rho_-(s)(k - k')}{\rho_+(1) |\bar{F}_L \setminus F^{(0)}|} \right] (2q_{L-1}) \\ &\leq (2\mu)^{-1} (2q_{L-1}). \end{aligned}$$

Again this inequality also holds when  $|\bar{F}_L \setminus F^{(k)}| = 0$ , and in such case (8) does not apply. This is because in this case  $Q(\mathbf{x}^{(k)}) \leq \min_{\mathbf{x}: \text{supp}(\mathbf{x}) \subset \bar{F}_L} Q(\mathbf{x}) \leq Q(\bar{\mathbf{x}}) + q_L$ . This finishes the induction. ■

The following lemma is a slightly stronger version of the theorem, which we can prove more easily by induction.

*Lemma A.5:* Consider the OMP algorithm. If there exist  $k$  and  $s$  such that  $|\bar{F} \cup F^{(k)}| \leq s$  and

$$k = \left\lceil 4|\bar{F} \setminus F^{(0)}| \frac{\rho_+(1)}{\rho_-(s)} \ln \frac{20\rho_+ (|\bar{F} \setminus F^{(0)}|)}{\rho_-(s)} \right\rceil,$$

then

$$Q(\mathbf{x}^{(k)}) \leq Q(\bar{\mathbf{x}}) + 2.5\epsilon_s(\bar{\mathbf{x}})^2/\rho_-(s). \quad (9)$$

*Proof:* We prove this result by induction on  $|\bar{F} \setminus F^{(0)}|$ . If  $|\bar{F} \setminus F^{(0)}| = 0$ , then the bound in (9) holds trivially because  $Q(\mathbf{x}^{(k)}) \leq Q(\mathbf{x}^{(0)}) \leq Q(\bar{\mathbf{x}})$ .

Assume that the claim holds with  $|\bar{F} \setminus F^{(0)}| \leq m-1$  for some  $m > 0$ . Now we consider the case of  $|\bar{F} \setminus F^{(0)}| = m$ . Without loss of generality, we assume for notational convenience that  $\bar{F} \setminus F^{(0)} = \{1, \dots, m\}$ , and  $|\bar{\mathbf{x}}_j|$  in  $\bar{F} \setminus F^{(0)}$  is arranged in descending order so that  $|\bar{\mathbf{x}}_1| \geq |\bar{\mathbf{x}}_2| \geq \dots \geq |\bar{\mathbf{x}}_m|$ . Let  $L$  be the smallest positive integer such that for all  $1 \leq \ell < L$ , we have

$$\sum_{i=2^{\ell-1}}^m \bar{\mathbf{x}}_i^2 < \mu \sum_{i=2^\ell}^m \bar{\mathbf{x}}_i^2,$$

but

$$\sum_{i=2^{L-1}}^m \bar{\mathbf{x}}_i^2 \geq \mu \sum_{i=2^L}^m \bar{\mathbf{x}}_i^2, \quad (10)$$

where  $\mu = 10\rho_+(m)/\rho_-(s)$ . We have  $L \leq \lfloor \log_2 m \rfloor + 1$  because the second inequality is automatically satisfied when  $L = \lfloor \log_2 m \rfloor + 1$  (the right hand side is zero in this case). Moreover, if the second inequality is always satisfied for all  $L \geq 1$ , then we can simply take  $L = 1$  (and ignore the first inequality).

We can now define

$$\bar{F}_\ell = (\bar{F} \cap F^{(0)}) \cup \{i : 1 \leq i \leq \min(m, 2^\ell - 1)\}$$

for  $\ell = 0, 1, 2, \dots, L$ .

Lemma A.1 implies that for  $\ell = 0, 1, \dots, L$ :

$$\min_{\mathbf{x} \in \bar{F}_\ell} Q(\mathbf{x}) \leq Q(\bar{\mathbf{x}}) + q_\ell,$$

$$q_\ell = 1.5\rho_+(m) \sum_{i=2^{\ell-1}}^m \bar{\mathbf{x}}_i^2 + 0.5\epsilon_s(\bar{\mathbf{x}})^2/\rho_+(m).$$

Moreover  $q_{\ell-1} \leq \mu q_\ell$  when  $\ell = 1, \dots, L-1$ . We can thus apply Lemma A.4 to conclude that when

$$\begin{aligned} k &= \sum_{j=1}^L \lceil (2^j - 1)(\rho_+(1)/\rho_-(s)) \ln(2\mu) \rceil \\ &\leq 2^{L+1}(\rho_+(1)/\rho_-(s)) \ln(2\mu) - 1, \end{aligned} \quad (11)$$

we have

$$\begin{aligned} &Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}}) \\ &\leq 1.5\rho_+(m) \sum_{i=2^L}^m \bar{\mathbf{x}}_i^2 + 1.5\mu^{-1}\rho_+(m) \sum_{i=2^{L-1}}^m \bar{\mathbf{x}}_i^2 \\ &\quad + 0.5(1 + \mu^{-1})\epsilon_s(\bar{\mathbf{x}})^2/\rho_+(m) \\ &\leq 3\mu^{-1}\rho_+(m) \sum_{i=2^{L-1}}^m \bar{\mathbf{x}}_i^2 + \frac{0.5}{\rho_+(m)}(1 + \mu^{-1})\epsilon_s(\bar{\mathbf{x}})^2, \end{aligned} \quad (12)$$

where (10) is used to derive the second inequality.

Now, if

$$2\mu^{-1}\rho_+(m) \sum_{i=2^{L-1}}^m \bar{\mathbf{x}}_i^2 \leq (1 + \mu^{-1})\epsilon_s(\bar{\mathbf{x}})^2/\rho_-(s), \quad (13)$$

then (12) implies that (9) holds automatically (since  $\mu \geq 10$ ), which finishes the induction. Therefore in the following, we only consider the case (13) does not hold, which implies that

$$2\mu^{-1}\rho_+(m) \sum_{i=2^{L-1}}^m \bar{\mathbf{x}}_i^2 > (1 + \mu^{-1})\epsilon_s(\bar{\mathbf{x}})^2/\rho_-(s).$$

Now Lemma A.2 implies that

$$\begin{aligned} &\rho_-(s) \|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\|_2^2 \\ &\leq 2(Q(\mathbf{x}^{(k)}) - Q(\bar{\mathbf{x}})) + \epsilon_s(\bar{\mathbf{x}})^2/\rho_-(s) \\ &\leq 6\mu^{-1}\rho_+(m) \sum_{i=2^{L-1}}^m \bar{\mathbf{x}}_i^2 + (2 + \mu^{-1})\epsilon_s(\bar{\mathbf{x}})^2/\rho_-(s) \\ &< 10\mu^{-1}\rho_+(m) \sum_{i=2^{L-1}}^m \bar{\mathbf{x}}_i^2 = \rho_-(s) \sum_{i=2^{L-1}}^m \bar{\mathbf{x}}_i^2. \end{aligned}$$

This implies that

$$\sum_{i=m-|\bar{F} \setminus F^{(k)}|+1}^m \bar{\mathbf{x}}_i^2 \leq \sum_{i \in \bar{F} \setminus F^{(k)}} \bar{\mathbf{x}}_i^2 \leq \|\mathbf{x}^{(k)} - \bar{\mathbf{x}}\|_2^2 < \sum_{i=2^{L-1}}^m \bar{\mathbf{x}}_i^2.$$

Therefore  $m - |\bar{F} \setminus F^{(k)}| + 1 > 2^{L-1}$ . That is,  $|\bar{F} \setminus F^{(k)}| \leq m - 2^{L-1}$ . It follows from the induction hypothesis that after another

$$\lceil 4(m - 2^{L-1})(\rho_+(1)/\rho_-(s)) \ln(2\mu) \rceil$$

OMP iterations, (9) holds. Therefore by combining this estimate with (11), we know that the total number of OMP iterations for (9) to hold (starting with  $F^{(0)}$ ) is no more than

$$\begin{aligned} &\lceil 4(m - 2^{L-1})(\rho_+(1)/\rho_-(s)) \ln(2\mu) \rceil \\ &\quad + 2^{L+1}(\rho_+(1)/\rho_-(s)) \ln(2\mu) - 1 \\ &\leq \lceil 4m(\rho_+(1)/\rho_-(s)) \ln(2\mu) \rceil. \end{aligned}$$

This finishes the induction step for the case  $|\bar{F} \setminus F^{(0)}| = m$ . ■

**Tong Zhang** Tong Zhang received a B.A. in mathematics and computer science from Cornell University in 1994 and a Ph.D. in Computer Science from Stanford University in 1998. After graduation, he worked at IBM T.J. Watson Research Center in Yorktown Heights, New York, and Yahoo Research in New York city. He is currently a professor of statistics at Rutgers University. His research interests include machine learning, algorithms for statistical computation, their mathematical analysis and applications.

## REFERENCES

- [1] P. Bechler and P. Wojtaszczyk. Error estimates for orthogonal matching pursuit and random dictionaries. *Constructive Approximation*, 2010. to appear.
- [2] T. Blumensath and M. E. Davies. Gradient pursuit for non-linear sparse signal modelling. In *European Signal Processing Conference (EUSIPCO)*, 2008.
- [3] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. on Information Theory*, 51:4203–4215, 2005.
- [4] M. A. Davenport and M. B. Wakin. Analysis of orthogonal matching pursuit using the restricted isometry property. *Information Theory, IEEE Transactions on*, 56(9):4395–4401, 2010.
- [5] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Info. Theory*, 52(1):6–18, 2006.
- [6] A. Fletcher and S. Rangan. Orthogonal matching pursuit from noisy random measurements: A new analysis. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 540–548. 2009.
- [7] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. Technical report, Rutgers University, January 2009. A short version appears in ICML’09. Available from <http://arxiv.org/abs/0903.3002>.
- [8] E. Liu and V. N. Temlyakov. Orthogonal super greedy algorithm and application in compressed sensing. Preprint, 2010.
- [9] E. Livshitz. On efficiency of orthogonal matching pursuit. Preprint, 2010.
- [10] D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2008.
- [11] D. Needell and R. Vershynin. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Foundations of Computational Mathematics*, 9(3):317–334, 2009.
- [12] H. Rauhut. On the impossibility of uniform sparse reconstruction using greedy methods. *Sampl. Theory Signal Image Process.*, 7(2):197–215, 2008.
- [13] S. Shalev-Shwartz, N. Srebro, and T. Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *Siam Journal on Optimization*, 20:2807–2832, 2010.
- [14] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Info. Theory*, 50(10):2231–2242, 2004.
- [15] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Info. Theory*, 53(12):4655–4666, 2007.
- [16] M. Warmuth, J. Liao, and G. Ratsch. Totally corrective boosting algorithms that maximize the margin. In *Proceedings of the 23rd international conference on Machine learning*, 2006.
- [17] C.-H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.
- [18] T. Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555–568, 2009.
- [19] T. Zhang. Some sharp performance bounds for least squares regression with  $L_1$  regularization. *Ann. Statist.*, 37(5A):2109–2144, 2009.